XÂY DỰNG THUẬT TOÁN HIỆU QUẢ CHO ĐỊNH GIÁ BẤT ĐỘNG SẢN QUẬN LONG BIÊN VÀ TỈNH MONTREAL

Nguyễn Hoàng Huy[1*], Phạm Văn Toàn[2], Hoàng Thị Thanh Giang[1]

[1]Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam
[2]Trường đại học Bách khoa Hà Nội

Email[*]: nhhuy@vnua.edu.vn

Ngày gửi bài: 04.12.2015 Ngày chấp nhận: 12.07.2016

TÓM TẮT

Phương pháp *LASSO* (Hastie *et al*., 2015) chỉnh hóa các hệ số *hồi quy tuyến tính* bằng cách thêm vào tiêu chuẩn bình phương tối tiểu một đại lượng phạt chuẩn ℓ_1. Gần đây, phương pháp này được sử dụng phổ biến để giải quyết các bài toán hồi quy số chiều cao trong các lĩnh vực thống kê, khai phá, học máy cho dữ liệu lớn. Trong bài báo này chúng tôi áp dụng phương pháp *LASSO* để chỉnh hóa các hệ số *hồi quy phi tuyến* cho bài toán định giá bất động sản. Định giá bất động sản thường chỉ dựa vào khoảng vài chục thuộc tính và rõ ràng mối liên hệ giữa giá bất động sản và các thuộc tính này không phải tuyến tính (Król, 2015), nên chúng tôi phải sử dụng mô hình phi tuyến. Khi đó số hệ số cần xác định trong mô hình này thường rất lớn, vì vậy chúng tôi áp dụng phương pháp *LASSO* để chỉnh hóa các hệ số này. Tuy nhiên phương pháp *LASSO* áp dụng như trên lại thường khá nhạy với tham số chỉnh hóa. Do đó chúng tôi đề xuất thuật toán *kết tập hồi quy phi tuyến LASSO* để cộng hưởng các hàm hồi quy *LASSO* yếu thành hàm hồi quy mạnh, có phương sai nhỏ hơn. Thuật toán này đã được đánh giá trên các tập dữ liệu giá bất động sản thu thập tại tỉnh Montreal, Canada (Noseworthy, 2014) và quận Long Biên, Hà Nội và cho kết quả chính xác hơn các thuật toán mới nhất đã được đưa ra.

Từ khóa: Giá bất động sản, hồi quy phi tuyến, hồi quy tuyến tính, phương pháp *LASSO*, kết tập hồi quy phi tuyến *LASSO*.

Building an Efficient Algorithm for Long Bien District and Montreal Real Estate Pricing

ABSTRACT

The *LASSO* method regularizes *linear regression* coefficients by adding a ℓ_1 norm penalty to the least square criterion. Recently, this method has been used very popularly to solve high dimensional regression problems in statistics, data mining, and machine learning for big data. In this paper, we applied the *LASSO* method to regularize *nonlinear regression* coefficients for the real estate pricing problem. Real estate pricing was often based on a few dozen features, and obviously the relationship between real estate prices and their features is nonlinear. Therefore in the present study we used a nonlinear model and applied *LASSO* method to regularize the coefficients. Because the performance of *LASSO* application is sensitive with regularization parameter, we proposed an *aggregation of LASSO nonlinear regression* combining weak *LASSO* regressions to produce a robust one which has smaller variance. This algorithm was evaluated on the real estate datasets collected in Montreal province, Canada (Noseworthy, 2014) and in Long Bien district of Hanoi and more accurate results than the state of the art algorithms were obtained.

Keywords: Real estate prices, linear regression, nonlinear regression, *LASSO* method, aggregation of *LASSO* nonlinear regression.

1. ĐẶT VẤN ĐỀ

Mỗi người chúng ta thường sẽ thực hiện giao dịch bất động sản ít nhất một lần trong đời. Số tiền dành cho mua nhà là không nhỏ, vì vậy việc người mua quan tâm không chỉ ở việc lựa chọn được một ngôi nhà ưng ý mà còn xem giá cả có hợp lý hay không. Việc đánh giá giá trị của một bất động sản dĩ nhiên không phải là một việc dễ dàng. Để đánh giá chính xác giá của một căn nhà, người ta không chỉ đòi hỏi một sự hiểu biết chuyên môn về thị trường bất động sản (một thị trường rất biến động) mà còn đòi hỏi một sự hiểu biết thật sự tường tận về bản thân các thuộc tính của bất động sản đó (Mu *et al.*, 2014). Những kiến thức này thường chỉ được lưu trữbởi các đại lý kinh doanh bất động sản. Nếu chúng ta có thể nắm bắt kiến thức này bằng cách thu thập dữ liệu, sử dụng các dữ liệu mở, tận dụng sự giúp sức của các thuật toán, chương trình máy tính, các kiến thức này trở nên dễ tiếp cận hơn với các người dân bình thường, giúp đưa ra quyết định mà không cần dựa vào chuyên gia vì không may vị chuyên gia đó có thể tư vấn theo chiều hướng có lợi cho họ.

Ước lượng giá bất động sản là một vấn đề hết sức quan trọng trong quy hoạch các thành phố lớn tại Việt Nam. Hiện nay, ở Việt nam chúng ta chủ yếu ước lượng giá bất động sản dựa trên các phương pháp truyền thống như phương pháp so sánh trực tiếp, chiết trừ, thu nhập, thặng dư, hệ số điều chỉnh. Các phương pháp này chủ yếu nhờ sự phân tích và can thiệp của nhân viên định giá nên rất khó tránh khỏi sai lầm do chủ quan hoặc không minh bạch (Quỳnh và cs., 2015). Ngoài các phương pháp truyền thống, trên thế giới đã và đang nghiên cứu và áp dụng rộng rãi các phương pháp có sử dụng đến các mô hình toán học để xác định giá trị bất động sản. Mới nhất là công trình (Król, 2015) sử dụng *mô hình hodenic* để mô hình hóa giá bất động sản ở Ba Lan. Một cách tổng quát, trong *mô hình hoderic*, hàm giá của bất động sản phụ thuộc vào các thuộc tính của nó như vị trí so với trung tâm, gần đường, gần các khu tiện ích, diện tích nhà, số phòng ngủ, số tầng,... Các mô hình để xác định hàm giá có thể là các mô hình đơn giản như mô hình tuyến tính hay các mô hình phức tạp hơn như mô hình mũ, mô hình logarit,...

Đã có một số nghiên cứu về việc xây dựng mô hình định giá bất động sản sử dụng các thuật toán học máy. Một trong số những nỗ lực đáng quan tâm đó là việc định giá bất động sản tại Montreal (Noseworthy *et al.*, 2014). Kết quả từ bài báo này rất ấn tượng và có ảnh hưởng đến cách lựa chọn các thuộc tính trong dữ liệu của chúng tôi. Nhóm tác giả đó đã sử dụng *hồi quy tuyến tính, hồi quy tuyến tính LASSO* và *K - láng giềng gần nhất*. Lần lượt các phương pháp cho trung bình sai số tuyệt đối chấp nhận được. Đây cũng là những phương pháp mới nhất áp dụng cho định giá bất động sản tỉnh Montreal. Những kết quả này như một sự đảm bảo, định hướng chúng tôi điều tra, khảo sát và xây dựng mô hình định giá bất động sản tại quận Long Biên. Tuy nhiên, không muốn lặp lại các kết quả đã được công bố trước đó và cuối cùng bị ràng buộc bởi tập dữ liệu đã có, chúng tôi lựa chọn việc khám phá và sử dụng các đặc điểm khác miêu tả và mô hình hóa giá của các ngôi nhà trong quận Long Biên.

Trong bài báo này chúng tôi phát triển thuật toán *kết tập hồi quy phi tuyến LASSO* để xây dựng mô hình định giá bất động sản tại quận Long Biên. Hiệu năng của thuật toán được đánh giá trên dữ liệu bất động sản chúng tôi thu thập được trên quận Long Biên. Hơn nữa, chúng tôi so sánh một cách chi tiết hơn thuật toán đó với những thuật toán mới nhất cho định giá bất động sản tại tỉnh Montreal (Noseworthy *et al.*, 2014). Đây là tập dữ liệu đã được công bố quốc tế rộng rãi.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP

2.1. Vật liệu nghiên cứu

2.1.1. Tập dữ liệu bất động sản quận Long Biên

Để thử nghiệm các thuật toán và mô hình đề xuất, chúng tôi sử dụng tập dữ liệu được chúng tôi điều tra trên địa bàn quận Long Biên, theo đề tài trọng điểm T2014 - 10 - 04 TĐ, tài

trợ bởi Học viện Nông nghiệp Việt Nam. Tập dữ liệu này bao gồm thông tin của 487 bất động sản, các thông tin này bao gồm: giá giao dịch, diện tích của khu đất, vị trí của khu đất chia theo quy định của Bộ Tài nguyên và Môi trường, độ rộng đường vào nhà, độ thuận tiện của lối vào nhà, khoảng cách đến trung tâm thành phố, khoảng cách đến trường học gần nhất, đánh giá chất lượng trường học, đánh giá chất lượng dịch vụ y tế, đánh giá trình trạng số đỏ, khoảng cách đến chợ gần nhất, khoảng cách đến trung tâm quận, độ rộng mặt tiền của thửa đất, tổng diện tích sàn của nhà, đặc điểm nhà,...

2.1.2. Tập dữ liệu bất động sản tỉnh Montreal

Trong bài báo này, chúng tôi đánh giá hiệu năng của các thuật toán học, mô hình định giá một cách chi tiết hơn trên tập dữ liệu bất động sản thu thập tại tỉnh Montreal. Đây là tập dữ liệu đã được công bố quốc tế. Tập dữ liệu mẫu này bao gồm các mô tả tiêu chuẩn của mỗi ngôi nhà cũng như số lượng các cơ sở hạ tầng trong vòng bán kính 3 km tính từ ngôi nhà đó.

Trong 9.717 mẫu dữ liệu thu thập được có những ngôi nhà không có đủ các thuộc tính. Rõ ràng các thuộc tính bị thiếu ảnh hưởng đến việc định giá của bất động sản đó. Noseworthy *et al.* (2014) đưa ra ba hướng tiếp cận để giải quyết vấn đề mất mát thông tin đó là: loại bỏ các bản ghi có các thuộc tính mất mát, dự đoán giá trị bị mất mát với phương pháp tối đa hóa kỳ vọng và thay giá trị bị mất với giá trị trung bình của các thuộc tính. Các tác giả đã chỉ ra rằng phương pháp bỏ đi các bản ghi bị mất mát là hiệu quả nhất trong xây dựng mô hình định giá. Khi đó tập dữ liệu bị rút gọn xuống còn chỉ *2.289* bản ghi. Trong bài báo này, tập dữ liệu rút gọn sẽ được sử dụng để đánh giá hiệu năng của các thuật toán học.

2.2. Phương pháp nghiên cứu

Hồi quy tuyến tính và *hồi quy tuyến tính LASSO* đã được áp dụng hiệu quả cho tập dữ liệu bất động sản tại tỉnh Montreal. Tuy nhiên giả thuyết giá bất động sản tuân theo mô hình tuyến tính rõ ràng không thỏa đáng (Król, 2014). Hơn nữa, *hồi quy tuyến tính LASSO* được đưa ra để giải quyết bài toán *hồi quy tuyến tính* cho dữ liệu thưa số chiều cao (số lượng thuộc tính lớn so với số bản ghi). Do vậy chỉ với vài chục thuộc tính thì giả thuyết các thuộc tính này thưa là thực sự không cần thiết (Noseworthy *et al.*, 2014). Hơn nữa, trong *hồi quy tuyến tính LASSO* vấn đề lựa chọn tham số chỉnh hóa tốt nhất không phải là công việc dễ dàng khi số bản ghi chỉ hàng trăm như trong dữ liệu bất động sản quận Long Biên. Trong bài báo này, chúng tôi lựa chọn một mô hình *hồi quy phi tuyến* thích hợp cho định giá bất động sản. Do số hệ số cần khớp lớn, chúng tôi áp dụng phương pháp *LASSO* để chỉnh hóa các hệ số này. Ở đây thay vì sử dụng các phương pháp lựa chọn tham số chỉnh hóa *LASSO* như *kiểm tra chéo*,... chúng tôi giới thiệu một phương pháp kết tập dựa trên nguyên lý *học tổ hợp* (*ensemble learning*) để kết hợp các hàm hồi quy *LASSO* yếu (chưa chính xác) thành một hàm hồi quy mạnh (chính xác hơn). Theo lý thuyết khái quát hóa làm sáng tỏ sự thành công của phương pháp *boosting* (một trong những phương pháp *học tổ hợp* điển hình) thì sự đa dạng, biến động của các hàm hồi quy *LASSO* khi qua các tham số chỉnh hóa khác nhau sẽ làm tăng hiệu năng của phương pháp kết tập. Mô hình hàm hồi quy sẽ được xây dựng trên tập dữ liệu huấn luyện và được đánh giá cuối cùng trên tập dữ liệu kiểm tra. Phương pháp *kiểm tra chéo* 5 phần đã được sử dụng để phân chia dữ liệu huấn luyện và kiểm tra. Dưới đây là mô tả cơ bản của thuật toán.

2.2.1. Hồi quy tuyến tính

Mô hình tuyến tính là một mô hình đơn giản và được sử dụng nhiều trong bài toán xác định giá bất động sản. Trong các nghiên cứu về giá bất động sản có sử dụng đến mô hình tuyến tính chúng ta có thể kể đến các nghiên cứu của (Christian el al., 2009; Richard, 2009). *Hồi quy tuyến tính* xác định một đường thẳng hay một mặt phẳng qua các điểm dữ liệu trong không gian thuộc tính. Giả sử giá của bất động sản là y và các thuộc tính ảnh hưởng đến giá của nó như diện tích, độ rộng mặt tiền, độ rộng đường vào nhà, tình trạng pháp lý của khu đất, tiện ích của khu dân cư (điều kiện vệ sinh, điều kiện

trường học, y tế), khoảng cách đến trung tâm phường, quận, thành phố... được lượng hóa và kí hiệu là $x_1, x_2, ..., x_p$. Ta cần xây dựng hàm giá của bất động sản là một hàm tuyến tính theo các biến trên, nghĩa là có dạng sau:

$$y = f(x_1, x_2, \ldots, x_p) = w_0 + \sum_{k=1}^{p} w_k x_k$$

Qua điều tra số liệu ta thu thập được *n* bộ số liệu và giả sử $y^i, x_1^i, x_2^i, ..., x_p^i,\ i = 1,\ 2, ..., n$ là các số liệu của bản ghi thứ *i*. Thông thườngta đi tìm các hệ số $w_k, k = 0, 1, 2, ..., p$ sao cho bình phương sai số là nhỏ nhất. Điều này dẫn đến việc giải một bài toán tối ưu như sau:

$$\min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(w_0 + \sum_{k=1}^{p} w_k x_k^i - y^i\right)^2\right\}$$

Đây là một bài toán tối ưu lồi, khả vi và không khó khăn để giải bài toán này bằng các công cụ khác nhau. Phương pháp hướng giảm thường được sử dụng để giải quyết vấn đề này. *Hồi quy tuyến tính* là một phương pháp hay không phải bởi vì nó là một phương pháp phổ biến được sử dụng trong các mô hình kinh tế mà còn bởi vì nó có một sự giải thích rất trực quan. Dựa trên độ lớn của các trọng số, chúng ta có thể thấy thuộc tính nào có tầm ảnh hưởng lớn đến giá trị của một ngôi nhà.

2.2.2. Mô hình phi tuyến LASSO

Thực tế thì mô hình *hồi quy tuyến tính* là đơn giản về phương pháp giải nhưng lại khó cho ra một sai số đủ tốt vì hàm giá có thể là một hàm số phi tuyến (Król, 2015). Sau rất nhiều khảo sát ban đầu cũng như tham khảo (Quỳnh và cs., 2015), chúng tôi đề xuất xấp xỉ căn bậc hai hàm giá bất động sản bằng một hàmbậc hai của các căn bậc hai các biến (thuộc tính).

$$\sqrt{y} = w_0 + \sum_{k=1}^{p} w_k x_k + \sum_{k=2}^{p}\sum_{l=1}^{k-1} w_{kl}\sqrt{x_k}\sqrt{x_l}$$

Khi đó hàm giá bất động sản được xác định bởi hàm hồi quy (1):

Với các giả thiết và điều kiện như trong phần *hồi quy tuyến tính* thì ta phải đi tìm các hệ số w_k, w_{kl} bằng phương pháp bình phương tối tiểu, nghĩa là giải bài toán tối ưu:

$$\min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(w_0 + \sum_{k=1}^{p} w_k x_k^i + \sum_{k=2}^{p}\sum_{l=1}^{k-1} w_{kl}\sqrt{x_k^i}\sqrt{x_l^i} - \sqrt{y^i}\right)^2\right\}$$

Mặc dù mô hình này khái quát hơn mô hình tuyến tính nhưng nó có nhược điểm là có nhiều tham số nên khi dung lượng mẫu không đủ lớn thì dễ dẫn đến hiện tượng học quá (Hastie *et al.*, 2009). Hiện tượng này dẫn đến sai số đo được trên dữ liệu huấn luyện nhỏ nhưng trên dữ liệu kiểm tra thì rất lớn. Có hai lý do lý giải cho hiện tượng này. Thứ nhất là khi sử dụng phương pháp bình phương tối tiểu thường có sai lệch thấp nhưng phương sai lớn và sự chính xác của dự đoán có thể được cải thiện bằng cách chỉnh hóa các hệ số hồi quy hoặc đặt một số hệ số bằng không. Bằng cách này, chúng ta có thể đưa thêm một vài sai lệch nhưng giảm phương sai của giá trị được dự đoán và do đó có thể cải thiện sự chính xác dự đoán toàn bộ (như trung bình sai số tuyệt đối). Lý do thứ hai cho sự giải thích được. Với số lượng lớn các hệ số, chúng ta thường xác định tập con nhỏ hơn các hệ số thực sự có nghĩa ảnh hưởng đến hàm hồi quy. Trong bài báo này chúng tôi sử dụng phương pháp *LASSO* để chỉnh hóa các hệ số của mô hình *hồi quy phi tuyến* trên. Phương pháp *LASSO* tìm các hệ số $w_k,\ w_{kl}$ bằng cách giải bài toán tối ưu (2).

Cận trên *t* là một kiểu "ngân sách": nó giới hạn tổng giá trị tuyệt đối của các hệ số cần ước lượng. Để thuận tiện bài toán *LASSO* thường được viết lại dưới dạng Lagrange với $\lambda \geq 0$. Do đối ngẫu Lagrange, có một tương ứng một - một giữa bài toán tối ưu có điều kiện ràng buộc (2) và dạng Lagrange (3).

$$y = f(x_1, x_2, \ldots, x_p) = \left(w_0 + \sum_{k=1}^{p} w_k x_k + \sum_{k=2}^{p}\sum_{l=1}^{k-1} w_{kl}\sqrt{x_k}\sqrt{x_l}\right)^2 \quad (1)$$

$$\min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(w_0+\sum_{k=1}^{p}w_k x_k^i+\sum_{k=2}^{p}\sum_{l=1}^{k\text{-}1}w_{kl}\sqrt{x_k^i}\sqrt{x_l^i}-\sqrt{y^i}\right)^2\right\} sao\,cho\sum_{k=1}^{p}|w_k|+\sum_{k=2}^{p}\sum_{l=1}^{k-1}|w_{kl}|\le t \quad (2)$$

$$\min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(w_0+\sum_{k=1}^{p}w_k x_k^i+\sum_{k=2}^{p}\sum_{l=1}^{k\text{-}1}w_{kl}\sqrt{x_k^i}\sqrt{x_l^i}-\sqrt{y^i}\right)^2+\lambda\left(\sum_{k=1}^{p}|w_k|+\sum_{k=2}^{p}\sum_{l=1}^{k-1}|w_{kl}|\right)\right\} \quad (3)$$

2.2.3. Kết tập hồi quy phi tuyến LASSO

Thuật toán *kết tập hồi qui phi tuyến LASSO* sẽ áp dụng mô hình *hồi quy phi tuyến* kết hợp với phương pháp *LASSO* như đã miêu tả ở trên. Tuy nhiên sai số của mô hình biến động theo sự lựa chọn tham số λ. Do đó trong bài báo này, chúng tôi đưa ra phương pháp khắc phục nhược điểm đó bằng cách kết hợp các hàm hồi quy này (tương ứng với các giá trị λ khác nhau). Thuật toán gồm các bước như sau:

Bước 1: Tìm các hệ số w_k^o, w_{kl}^o từ phương trình (3) tương ứng với giá trị khởi tạo tham số chỉnh hóa $\lambda_0 = 0$, ước lượngtrung bình sai số tuyệt đối e^o của dữ liệu huấn luyện

Bước 2: Tính $\lambda_m = \lambda_0 + m\nabla\lambda$ và tìm các hệ số w_k^m, w_{kl}^m từ phương trình (3) tương ứng với giá trị $\lambda = \lambda_m$, ước lượng trung bình sai số tuyệt đối e^m của dữ liệu huấn luyện ($\nabla\lambda$=0,005)

Lặp lại bước 2 cho m=1,2,…cho đến khi $e^m \ge e_0 + e^\nabla$ (e^∇=5.000), khi đó ở bước cuối cùng ta được $m=M$. Các mô hình *hồi quy phi tuyến LASSO* này được kết tập lại hình thành một tổ hợp *hồi quy phi tuyến*:

$$w_k = \frac{1}{M+1}\sum_{m=0}^{M} w_k^m, w_{kl} = \frac{1}{M+1}\sum_{m=0}^{M} w_{kl}^m$$

Những hệ số này sẽ được dùng để xây dựng mô hình hồi quy cuối cùng cho định giá bất động sản, hàm giá bất động sản được cho bởi công thức (1). Phương pháp này không chỉ thực hiện sự lựa chọn các hệ số có nghĩa một cách tự động mà còn làm giảm phương sai để cải thiện khả năng khái quát hóa của mô hình.

3. KẾT QUẢ VÀ THẢO LUẬN

Hiệu năng của các mô hình *hồi quy tuyến tính, hồi quy phi tuyến* có và không áp dụng phương pháp chỉnh hóa *LASSO* và *kết tập hồi quy phi tuyến LASSO* được so sánh trên tập dữ liệu bất động sản tỉnh Montreal. Noseworthy *et al.* (2014) đã chỉ *hồi quy tuyến tính* có hiệu năng tương đương với *hồi quy tuyến tính LASSO* và các tác giả cũng chỉ ra đây là những phương pháp định giá bất động sản thích hợp, cho kết quả tốt trên tập dữ liệu thu thập tại tỉnh Montreal. Với những kết quả thực nghiệm chỉ ra dưới đây chúng ta có thể thấy *kết tập hồi quy phi tuyến LASSO* cho sai số tương đối chính xác hơn khoảng *2%* so với những phương pháp kể trên (giá trung bình của các bất động sản tỉnh Montreal thu thập được là *312.380 $*).

3.1. Hồi quy tuyến tính và hồi quy tuyến tính LASSO

Bảng 1 cho ta kết quả chi tiết của trung bình sai số của phương pháp *hồi quy tuyến tính LASSO* qua các giá trị λ = 0; 1; 5; 10; 100; 1.000. Với λ = 0 *hồi quy tuyến tính LASSO* trở thành*hồi quy tuyến tính*. Ta có thể thấy trung bình sai số tuyệt đối ổn định trừ phi λ nhận giá trị rất lớn cỡ hàng nghìn. Hiệu suất tốt nhất của *hồi quy tuyến tính LASSO* trên tập dữ liệu bất động sản tỉnh Montreal là ứng với λ=100, nó mang lại trung bình sai số tuyệt đối là *46.557 $*.

3.2. Hồi quy phi tuyến LASSO và kết tập hồi quy phi tuyến LASSO

Bảng 2 cho ta kết quả chi tiết của trung bình sai số tuyệt đối của *hồi quy phi tuyến LASSO* đã được xác định cụ thể trong phần 3 trên tập dữ liệu bất động sản tỉnh Montreal. Với

Bảng 1. Kết quả trung bình sai số tuyệt đối (trên dữ liệu kiểm tra, tỉnh Montreal) tương ứng với các giá trị của tham số chỉnh hóa λ của hồi quy tuyến tính LASSO

Hồi quy tuyến tính LASSO	λ = 0	λ = 1,0	λ = 5,0	λ = 10	λ = 100	λ = 1.000
Sai số	*46.677*	*46.676*	*46.668*	*46.654*	*46.557*	*47.383*

Bảng 2. Kết quả trung bình sai số tuyệt đối tương ứng với các giá trị của tham số chỉnh hóa λ của hồi quy phi tuyến LASSO trên tập dữ liệu huấn luyện và kiểm tra, tỉnh Montreal

Hồi quy phi tuyến LASSO	λ = 0	λ = 1	λ = 5	λ = 10	λ = 100	λ = 1.000
Trên dữ liệu huấn luyện	31.749	40.036	43.652	47.840	80.028	80.028
Trên dữ liệu kiểm tra	52.828	43.164	46.502	51.686	86.664	86.664

λ = 0 thì mô hình này trở thành mô hình *hồi quy phi tuyến* cũng được miêu tả cụ thể trong phần 3. Ta có thể thấy trung bình sai số tuyệt đối của mô hình phi tuyến khá nhỏ cho dư liệu huấn luyện (*31.749 $*) nhưng khá lớn cho dữ liệu kiểm tra (*52.828 $*). Còn sai số trung bình tuyệt đối của *hồi quy phi tuyến LASSO* trên dữ liệu kiểm tra biến động nhiều khi chạy qua các giá trị λ = 0; 1; 5; 10; 100; 1.000. Có nhiều giá trị λ cho trung bình sai số tuyệt đối nhỏ hơn so với mô hình phi tuyến không áp dụng phương pháp chỉnh hóa *LASSO*, ngược lại cũng có nhiều giá trị λ cho kết quả lớn hơn. Điều này có thể lý giải được bởi trong mô hình này số lượng các hệ số cần xác định là khá lớn lên đến *780* hệ số, tương ứng với *39* thuộc tính.

Chúng tôi áp dụng phương pháp kết hợp *hồi quy phi tuyến LASSO* cho tập dữ liệu bất động sản tỉnh Montreal. Chúng tôi khởi tạo giá trị tham số chỉnh hóa $\lambda_0 = 0$, bước nhảy tham số chỉnh hóa $\nabla\lambda = 0{,}005$ và ngưỡng độ chệnh trung bình sai số tuyệt đối là $e^{\nabla} = 5.000$. Trung bình sai số tuyệt đối của thuật toán*kết tập hồi quy phi tuyến LASSO* là *40.250 $*, nghĩa là sai số tương đối là *12,88%*.

Chúng tôi cũng đánh giá hiệu năng của thuật toán*kết tập hồi quy phi tuyến LASSO* với dữ liệu giá đất do chúng tôi thu thập tại quận Long Biên. Dữ liệu thô ban đầu gồm 50 thuộc tính và giá của bất động sản chuyển nhượng. Tuy nhiên dữ liệu này chứa nhiều thuộc tính bị mất thông tin. Chúng tôi loại những thuộc tính mất nhiều thông tin và bổ xung thêm các thuộc tính khai thác được từ Google Maps APIs được *41* thuộc tính, tương ứng với nó có *178* bản ghi chứa đầy đủ thông tin của *41* thuộc tính đã chọn. Kết quả hơi thất vọng khi sai số tương đối của thuật toán *kết tập hồi quy phi tuyến LASSO* chỉ đạt được trên dữ liệu kiểm tra *26,48%*.

3.3. Thảo luận

Các kết quả định giá bất động sản quận Long Biên không như mong đợi. Công trình (Noseworthy *et al.*, 2014) đã khiến chúng tôi hi vọng rằng chúng tôi có thể đạt được kết quả tượng tự. Có thể việc sử dụng một tập các thuộc tính riêng biệt là lý do tại sao trung bình sai số tuyệt đối thu được trong thực nghiệm của chúng tôi không thể so sánh với kết quả thu được trên tập dữ liệu bất động sản của tỉnh Montreal. Tuy nhiên các kết quả không thể so sánh một cách trực tiếp bởi vì vốn dĩ các thuộc tính trong tập dữ liệu của tỉnh Montreal và quận Long Biên là khác nhau. Hơn nữa tập dữ liệu về bất động sản quận Long Biên sau khi loại bỏ nhiều thuộc tính có thể chưa bao hàm đầy đủ các thông tin cần thiết cho việc định giá bất động sản. Hơn nữa, phần lớn các dữ liệu đều được thu thập từ các chủ bất động sản. Theo trực giác đáng lẽ các ngôi nhà gần nhau nếu có các thuộc tính tương tự nhau thì giá thành của chúng cũng phải tương tự nhau tuy nhiên trong tập dữ liệu này đôi lúc không phải vậy. Thực tế là các chủ căn nhà đều có xu hướng đánh giá rất chủ quan ngôi

nhà của mình. Tuy nhiên cũng có một số thành quả thu được từ việc thử nghiệm các thuật toán này. Quan trọng nhất là việc xây dựng thành công thuật toán định giá bất động sản trên tập dữ liệu bất động sản đã được công bố quốc tế của tỉnh Montreal. Những khảo sát của chúng tôi đã chỉ ra thuật toán *kết tập hồi quy phi tuyến LASSO* là tốt hơn các thuật toán mới nhất cho tập dữ liệu bất động sản tỉnh Montreal (Noseworthy *et al.*, *2014*) và cho sai số tương đối chỉ là *12,88%*. Đối với các mô hình tuyến tính, hiệu năng của chúng bị giảm có thể giải thích do sự phi tuyến tính của hàm giá bất động sản. Bởi vì thực sự thị trường nhà ở vốn là một thị trường vô cùng phức tạp, trên thực tế là không một ai có thể hiểu về nó thật sự thấu đáo.

4. KẾT LUẬN

Rõ ràng mô hình phi tuyến được lựa chọn cho phép chúng ta xây dựng mô hình dữ liệu bất động sản khái quát hóa hơn (Król, 2015). Tuy nhiên với số lượng lớn hệ số cần xác định của mô hình, lên đến *780* trong khi dữ liệu huấn luyện của mỗi phần trong kiểm tra chéo *5* phần chỉ là *1832* bản ghi như trong tập dữ liệu bất động sản tỉnh Montreal, nên việc học mô hình này thường dẫn đến hiện tượng học quá (Hastie *et al.*, 2009). Hiện tượng này thể hiện ở bảng 2 khi trung bình sai số tuyệt đối trên dữ liệu huấn luyện nhỏ nhưng trên dữ liệu kiểm tra lớn. Để khắc phục nhược điểm này chúng tôi áp dụng phương pháp *LASSO* thường sử dụng cho các mô hình *hồi quy tuyến tính* số chiều lớn để chỉnh hóa các hệ số khớp với mô hình. Tuy nhiên, trung bình sai số tuyệt đối khi đó biến động rất lớn khi tham số chỉnh hóa thay đổi. Do đó chúng tôi đưa ra thuật toán *kết tập hồi quy phi tuyến LASSO* dựa trên nguyên lý học tổ hợp để kết hợp các mô hình trên lại thành mô hình hồi quy hiệu quả hơn. Kết quả thực nghiệm chỉ ra phương pháp được đưa ra cho trung bình sai số tương đối chính xác hơn các thuật toán mới nhất cho dữ liệu bất động sản tỉnh Montreal khoảng *2%* (Noseworthy *et al.*, 2014). Tuy nhiên khi áp dụng thuật toán này cho tập dữ liệu bất động sản quận Long Biên thì hiệu quả không được như mong đợi. Có thể điều này là do sự đánh giá chủ quan của các chủ bất động sản khi được chúng tôi khảo sát, thu thập số liệu.

TÀI LIỆU THAM KHẢO

Christian G., Laferrère A. (2009). Managing hedonic housing price indexes: The French experience, Journal of Housing Economics, 18: 206 - 213.

Hastie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer.

Hastie T., Tibshirani R., Wainwright M. (2015). Statistical Learning with Sparsity, The Lasso and Generalizations, CRC Press.

Król A. (2015). Application of Hedonic Methods in Modelling Real Estate Prices in Poland, Data Science, Learning by Latent Structures, and Knowledge Discovery, pp. 501 - 511.

Mu J., Wu F., and Zhang A. (2014). Housing Value Forecasting Based on Machine Learning Methods, Abstract and Applied Analysis, 7 p. doi:10.1155/2014/648047

Noseworthy M., Schiazza B. L. (2014). Montreal Real Estate Pricing, Technical Report, McGillUniversity, Website: http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission 89.pdf.

Richard J. C. (2009). The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District, The Review of Regional Studies, 39(1): 9 - 22.

Trần Đức Quỳnh, Bùi Nguyên Hạnh (2015). Mô hình Hedonic và phần mềm cho bài toán xác định giá đất, các yếu tố ảnh hưởng đến giá đất. Tạp chí Khoa học và Phát triển, 13(6): 989 - 998.